

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/115515>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

# Cooperation in Public Goods Games Predicts Behavior in Incentive-Matched Binary Dilemmas: Evidence for Stable Pro-Sociality

## Abstract

We report the results of an experiment in which subjects completed second mover public goods game tasks and second mover binary social dilemma tasks. Each task was completed under three different incentive structures which were matched across tasks. The use of non-linear incentive structures, along with a novel categorization method, allowed us to identify behavioral subtypes that cannot be distinguished using conventional linear incentive structures. We also examined how well behavior could be predicted across tasks. Subjects' average conditional cooperation levels showed significant cross-task predictability and stability. However, almost a third of responses (28%) demonstrated unambiguous preference reversals across tasks. We argue that pro-sociality is best described as an individual-level trait, similar to risk aversion in choice under risk. (JEL Classifications: C7 C91 H41)

**Keywords:** Public Goods; Social Dilemmas; Cross-Task Prediction; Prisoner's Dilemma; Stag Hunt; Cooperation

# 1.Introduction

A key development in the study of cooperation has been the identification of behavioral subtypes. It has long been possible to categorize individuals as “cooperators” or “defectors” in terms of their performance on simple binary dilemma (BD) games such as the prisoner’s dilemma, where subjects explicitly choose between cooperation and defection within a standard game matrix. More recently, better than binary classification of individuals has been made possible by the use of public goods (PG) games which reveal the responsiveness of an individual’s contributions to the public good as a function of the contributions made by others to the same public good (see, e.g., Fischbacher, Gächter & Fehr 2001)<sup>1</sup>.

Two key questions are addressed here. The first concerns whether behavioral subtyping based on responses in one task can predict behavior on different tasks (e.g., whether cooperativeness on PG games predicts cooperativeness on BD games). Is “conditional cooperativeness” akin to a personality trait, i.e., a stable characteristic of an individual that governs their behavior across a wide range of contexts? Or do individuals apply such different strategies across tasks that it is not possible to use their behavioral subtype in a single task to predict their behavior in other tasks? A behavioral subtyping that does not predict performance across economic games within the laboratory is unlikely to predict behavior outside the laboratory. Furthermore, comparisons between more closely controlled tasks may help to identify why results vary so substantially between different tests of lab to field generalizability (Camerer, 2011).

The second question is whether there are additional, or alternative, behavioral subtypes that are only apparent in PG games with non-linear incentive structures. Prior investigations have largely used the linear PG game. However, this approach is limited in the number of strategies it can identify because quite differently motivated strategies can produce identical patterns of behavior. Specifically, “free-riders” are defined as individuals who contribute zero to the public pot regardless of the amount being contributed by other players. These individuals are often interpreted as payoff-maximizers, because the dominant strategy for a self-interested individual in the linear PG game is to contribute nothing regardless of others’ contributions<sup>2</sup>. However, zero contribution could alternatively reflect strategies of non-investment or non-engagement with the market. Similarly, it is unclear

---

<sup>1</sup> For reviews, see, e.g., Ledyard (1994) and Chaudhuri (2011).

<sup>2</sup> An additional point contributed to the public pot gives a return less than that from keeping it in the private account, regardless of the other players’ contributions.

whether conditional cooperators imitate other players or whether they base their decisions on some mix of factors such as equality of contribution, equality of outcome, and total group payout.

To address our two key questions we develop a novel within-subjects methodology in which individuals complete PG and BD tasks with the same three incentive structures being used in both tasks. Prior studies on the cross-task stability of social preferences have led to mixed conclusions. This is particularly true when comparing pro-social behavior in laboratory tasks to behavior in the real world. Some studies find significant correlations (Camerer, 2011; Dai, Galeotti, & Villeval, 2016; Karlan, 2005; Normann, Requate, & Waichman, 2014), but others report remarkably robust null effects across a wide range of tasks and measures (Carpenter & Seki, 2011; Galizzi & Navarro-Martinez, 2018; Stoop, Noussair, & van Soest, 2012). In many studies that do report a significant relationship, the correlation is often noticeably smaller than test-retest correlations when an individual completes the same task multiple times. Overall, the results suggest that many unidentified factors affect responses and measured pro-sociality (Fehr & Leibbrandt, 2011; Lamba & Mace, 2011).

However, it is not straightforward to identify specific features of study designs that could explain why some studies show an effect whilst others do not. This is largely because the many differences between the tasks used in the various studies make it impossible to determine the effect of each. Consider, for example, Stoop et al., (2012) which compares behavior in cooperation tasks in an abstract lab setting to behavior in comparable field studies at a recreational fishing pond. These authors find cooperation in the lab, but virtually none in the field, and convincingly demonstrate that this contrast is driven by task differences rather than population differences. However, identifying the mechanism underpinning the task effect is impossible, due to the many differences which could account for (or significantly contribute towards) the difference in cooperation. For example, social considerations (real or perceived) may have been more salient in the field because a subject can more immediately see who else is fishing in the pond, with potential perceived reputational costs. There are likely to be diminishing returns to catching fish, since subjects were required to take their catch home or dispose of it. Importantly for our own design, it is difficult to identify how well the incentives matched across tasks, since the utility of catching a fish is likely to vary substantially between subjects in a way that is not true for the money received in financial tasks. Stoop et al. made an impressive attempt to control for, or address, as many potential differences as possible. However, in this and other field studies, the differences in framing and incentives will be multi-faceted, and all but impossible to define comprehensively.

Although we focused on the example of Stoop et al., (2012) similar issues apply more generally across laboratory and field experiments. The potential for real or perceived social pressures and reputational effects will often differ between studies, or between tasks within a study. The immediacy or depth of social interactions will typically be different in computerized tasks compared to physical or verbal interactions in tasks in the field. Even the physical effort required to complete the tasks may differ. Undoubtedly, studies in the field generate insights that are valuable in their specific contexts, but due to the proliferation of differences discussed above, trying to identify the common causes of behavioral patterns will at best be a long and inefficient process. The potential variation in incentives, framing, and subject perceptions are simply too large. To address this, others have used entirely laboratory-based studies. Even in these studies, however, the evidence for cross-task predictability is mixed.

For example, Blanco et al. (2011) examined inequity aversion (Fehr & Schmidt, 1999) using four different tasks: an ultimatum game, a PG game, a dictator game, and a sequential prisoner's dilemma game. They found that the majority of players did not behave consistently across the different games in the way predicted by the inequity aversion model. There were, however, correlations across games — for example, individuals' sequential prisoner's dilemma second mover-decisions were correlated with their offers in the ultimatum game.

In a detailed analysis of the evidence for the inequity aversion model, Binmore and Shaked (2010) concluded that the cross-task predictive power of the model had not been established; i.e. it is not clear that parameters estimated from behavior on one task can be used to predict behavior on another (but for discussion, see Binmore & Shaked, 2010; Gintis, 2010). Even if payoff structures are identical, the level of cooperativeness may differ. For example, a number of studies have found greater cooperativeness in public goods games than in common pool resource games<sup>3</sup>, despite suggestions that the games are strategically equivalent (Ledyard & Palfrey, 1995). In contrast, other studies have found that overall levels of cooperation over repeated games are qualitatively similar when payoffs are equivalent across these tasks (Apesteguia & Maier-Rigaud, 2006), with results depending on experimental parameters (Kingsley & Liu, 2014). However, such studies have not examined the strategies employed by the same subjects in different tasks. This is particularly important due to the strong evidence of large individual differences in cooperation and strong behavioral subgrouping (Fischbacher, Gächter, & Fehr, 2001).

---

<sup>3</sup> Perhaps because of “warm glow” effects (Andreoni, 1995).

Here, we present a strategy that addresses all of the issues outlined above. We present a within-subject methodology that maintains the same financial incentive structure across two different lab tasks. This allows us to examine the extent to which cooperation can be generalized across different task frames and incentive structures. By allowing such a high level of control, this paradigm also presents a starting point from which the effect of specific experimental manipulations can be unambiguously measured.

We directly assess cross-task predictability and the effect of incentive structure. Subjects completed an experiment with two sections — a PG section and a BD section, and within these sections completed first and second mover responses. We focus on the second mover responses. For the PG games we adopted the strategy game methodology (“P-game”) used in Fischbacher, Gächter and Fehr (2001; see also Selten 1967). Subjects stated their conditional contribution in response to all possible integer values of the mean of others’ contributions (0 – 20). Subsequently they provided their first mover contribution (as in a “C-game”). The PG section included three versions of the game, each with a different incentive structure. The same three incentive structures were included in the BD section, as well as a fourth incentive structure used to test for violations of dominance. The incentive structures were chosen to correspond to three common binary dilemmas: prisoner’s dilemma, stag hunt and hawk-dove. In the PG task this correspondence was achieved by transforming the group summed contributions into the shared payoff from the public good according to either a linear, convex, or concave function (see below for details). The use of quadratic functions (following Isaac & Walker, 1988, Keser, 1996, Sefton & Steinberg, 1996, and others) also allowed us to examine whether there are subtypes of individuals that cannot be identified using only the linear structure. When responding to the BD, subjects answered both as first mover and second mover in all incentive structures. In the second mover case, subjects were told the other player(s) had chosen to cooperate. Our experiment exploits the equivalence between incentives in these second mover choices and in the PG strategy task.

Our principal findings are as follows. First, using a novel categorization method we grouped subjects into three behavioral subtypes in the PG game: 69% of subjects were conditional cooperators; 11% were payoff-maximizers, and 20% were non-contributors. Second, both payoff-maximizers and conditional cooperators were sensitive to the incentive structure of the PG game, although the effect was by definition smaller for those who were categorized as conditional cooperators. Only 9% of subjects exhibited a pure matching strategy (i.e. contributed the same amount as other players, regardless of the incentive structure). We also found evidence for cross-task stability: Cooperativeness in binary social dilemmas was significantly predicted by parameters describing an individual’s behavior in the PG games. Furthermore, this predictive accuracy was independent of incentive structure,

with predictive power no better within an incentive structure than between. We interpret this as evidence for pro-sociality as a stable trait that influences responses under all incentive structures. However, subjects also demonstrated variability across tasks. A substantial minority of individuals exhibited preference reversals between PG and BD, even when incentive structures were held constant (e.g. an individual cooperated in a BD but contributed little or nothing in the equivalent PG).

The remainder of the paper is structured as follows. In section 1.1, we describe the incentive structures and show how they can be varied in PG games to be made equivalent to the incentive structures of prisoner's dilemma, stag hunt, and hawk-dove games, while also allowing us to distinguish between the different possible strategies being used by non-investors and by conditional cooperators. The experimental design is described in section 2, and results are reported in section 3. Section 4 concludes.

## 1.1 Incentive Structures

As set out by Ledyard (1994), any public goods game environment can be described as follows. Each agent,  $i = 1, \dots, N$  in the group holds an endowment,  $z^i$ . The PG is produced from contributions out of this endowment, according to some function  $y = g(t)$  where  $t$  represents the part of the initial endowments that group members contribute. The rest ( $x^i$ ) is retained, so that  $t = \sum_{i=1}^N (z^i - x^i)$ . The group's outcome is then  $a = (y, x^1, \dots, x^N)$ . Each individual is assumed to derive utility from the PG and from their own private good, according to some function  $U^i(y, x^i)$ . The total payoff for each participant is typically given by  $x^i + \frac{y}{N}$ .

Some of these parameters were fixed for this study.  $N = 4$  in all PG games, and  $z_i = 20$  for all participants in each game. The functional form taken by  $g(t)$  is specified as  $g(t) = (\alpha t)^\beta$  where  $\alpha$  and  $\beta$  are varied to generate the linear, convex and concave incentives. In the linear form of the PG game,  $\beta = 1$  to generate linearity, and  $\alpha > 1$  to provide the tension between private and social incentives<sup>4</sup>. To generate convexity  $\beta > 1$ ; and for concavity,  $\beta < 1$ .

These incentive structures generate different second mover best-response functions under the assumption of payoff maximization. The payoff function for an individual can be summarized as

---

<sup>4</sup> That is, while the dominant strategy (and therefore the Nash Equilibrium) is to contribute no points to the PG, the Pareto Efficient outcome is for everybody to contribute all twenty points.

$$\pi^i = 20 - t^i + \frac{1}{4} \left( \alpha (T^J + t^i) \right)^\beta \quad (1)$$

$T^J$  is the sum of the other 3 players' contributions. Differentiating equation (1) with respect to  $t^i$  gives the result that contributing an additional point to the public pot increases one's own payoff as long as

$$1 < \frac{\alpha\beta}{4} \left( \alpha (T^J + t^i) \right)^{\beta-1} \quad (2)$$

Where  $\beta = 1$ , the condition in (2) simplifies to contribution whenever  $1 < \frac{\alpha}{4}$ , a fixed condition that, with  $\alpha = 1.4$ , is never satisfied. This generates the boundary solution where zero contributions are always the best response to any level of others' contribution.

In the convex case with  $\beta > 1$ , the benefit from contributing an additional point increases in both  $T^J$  and  $t^i$ . The more that has been contributed, the more valuable is an additional contribution. Substituting in the numbers used in this study, with  $\alpha = 0.07$  and  $\beta = 3$ , the threshold amount in the pot for which the value of an additional contribution exceeds 1 is 62.35 points.

In the concave case with  $\beta < 1$ , the benefit from contributing an additional point decreases in both  $T^J$  and  $t^i$ . The more that has been contributed, the less valuable is an additional contribution. The concave case employed in this study uses  $\alpha = 800$  and  $\beta = 0.4$ , and an additional point contributed to the public pot generates additional private return only when the total contributions are zero or 1 (the threshold is 1.85). Therefore, except when others contribute nothing, there is no payoff maximizing incentive to contribute to the PG in the concave case. The workings are provided in Appendix A.

So far, the focus has been on the PG incentive structures. But in order to explore the consistency of behavior across contexts, it is necessary to create scenarios that are equivalent in terms of their payoffs, but different in their framing. To do this, we make use of a little-recognized theoretical observation. Specifically, variants of the standard PG and BD tasks can be created such that the payoff structures are equivalent across the two task formats at particular levels of PG contribution (see Kollock, 1998, for a discussion along similar lines). Any strategy that relies solely on the payoffs (of the respondent and/or of the other players) will therefore lead to the same levels of cooperation across task formats.

To understand how the payoff-equivalence is constructed, first consider the one-shot linear PG task ("C-game") with a total of four players. In each round,  $t^i \in [0, 20]$ , so the decision maker contributes any number of points between 0 and 20, knowing that the



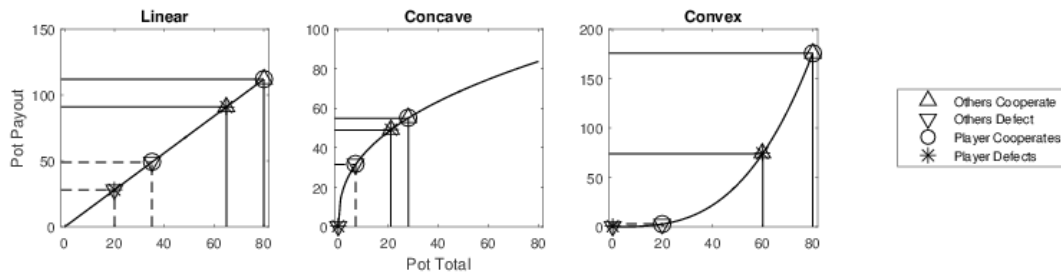
average of others' contributions will also be between 0 and 20. Consider instead a case where the decision maker's choice set is limited to two possible contributions; e.g.  $t^i \in \{5, 20\}$ . If all players face the same restricted choice<sup>5</sup>, the PG game reduces to a BD between cooperating (by contributing 20) and defecting (by contributing 5). This is illustrated in the leftmost panel of Figure 1.<sup>6</sup> The points marked on the pot payout functions show the possible combinations of cooperation and defection when the PG game is reduced to the  $\{5, 20\}$  binary dilemma described above. For example, the leftmost highlighted point shows the result of the Player and the Others all defecting, resulting in a pot of 20 and a pot payout of 28. The rightmost highlighted point shows the result of Player and Others all cooperating, with pot size 80 and pot payout 112. The four highlighted payoffs from these cooperate-defect choices form the prisoner's dilemma shown in panel a of Figure 2 (for a 2 player version) and Figure 3 (for a 4 player version). We are not the first to notice this equivalence, which is clearly set out in Hauert and Szabo (2003) and discussed in some detail by Conybeare (1984), but we are the first to use the equivalence to test consistency of behavior within-subjects across task frames.

Similar equivalences can be constructed for stag hunt games and for hawk-dove games. In a PG game where the transformation determining the public good payoff is convex in summed contributions, the benefit from contributing an additional point to the public pot increases in the average of others' contributions. When reduced to a binary choice between cooperation (contribute all) and defection (contribute nothing), the choice is equivalent to the stag hunt dilemma in panel c of Figures 2 and 3. This is illustrated in the context of the public goods game in the rightmost panel of Figure 1. If the payoff transformation is concave, as in the middle panel of Figure 1, the benefit of contributing an additional point to the public pot declines with the average of others' contributions. This can be reduced to the hawk-dove BD shown in panel b of Figures 2 and 3, where the best response is to defect when the other player(s) cooperate, but cooperate when the other(s) defect.

FIGURE 1

<sup>5</sup> The methods section details how a group would decide.

<sup>6</sup> The incentive structures and payoffs shown in this figure are those used in our experiment.



**Figure 1** Three Incentive Structures in Public Goods and Binary Dilemmas. The points marked on the pot payout functions show the possible combinations of cooperation and defection when the PG game is reduced to the binary dilemmas.

One difference between the PG game and the classic BD games is the number of other players. In the BD the decision maker typically faces just one other player instead of three others. To account for this, our design features 2-player and 4-player versions of the binary dilemmas. Table 1 and Table 2 give the payoffs for the 2-player and 4-player BD games, respectively.

**FIGURE 2**

Prisoner's Dilemma (linear)			Stag Hunt (convex)			Hawk Dove (concave)		
Player	Opponent		Player	Opponent		Player	Opponent	
	Cooperate	Defect		Cooperate	Defect		Cooperate	Defect
	Cooperate	Defect		Cooperate	Defect		Cooperate	Defect
	28 , 28	17.5 , 32.5		88 , 88	11 , 51		31.5 , 31.5	24 , 44
	32.5 , 17.5	22 , 22		51 , 11	40 , 40		44 , 24	20 , 20

**Figure 2** Payoff matrices for all binary dilemmas with one other player.

**FIGURE 3**

Prisoner's Dilemma (linear)			Stag Hunt (convex)			Hawk Dove (concave)		
Player	Opponents		Player	Opponents		Player	Opponents	
	Cooperate	Defect		Cooperate	Defect		Cooperate	Defect
	Cooperate	Defect		Cooperate	Defect		Cooperate	Defect
	28 , 28	12.25 , 27.5		44 , 44	1 , 21		26.5 , 26.5	21 , 28
	37.75 , 22.75	22 , 22		28.5 , 18.5	20 , 20		32 , 25	20 , 20

**Figure 3** Payoff matrices for binary dilemmas with three other players.

The Nash equilibria for the binary dilemmas with two players are as follows. The prisoner's dilemma has a single Nash Equilibrium (NE) of {D,D} despite the efficient outcome being {C,C}. This captures the private – social tradeoff and reflects the structure of the linear PG game. The hawk-dove has two pure strategies NE, {C,D} and {D,C}. That is, one should

defect if the other cooperates and vice versa. This logic is reflected in the concave PG game. There also exist two pure strategies NE for the stag hunt:  $\{C,C\}$  and  $\{D,D\}$ , reflecting the increasing returns to contributions in the convex incentive structure of the PG game.

Turning to the 4 player versions, the structure is that a single player is facing a group of three others who will vote on their preferred outcome. The NE are equivalent to those for 2-players for the concave and convex versions. However, it is not possible to select values in the prisoner's dilemma (linear case) with 3 other players that produce the incompatibility between payoff maximization and socially optimal outcomes, since the requirement that the other players divide the payoff by 3 precludes this. Therefore the NE for the 4 player version of the prisoner's dilemma is  $\{D,C\}$ . The 'group of others' has a dominant strategy to cooperate regardless of the single player's behavior. While this clearly changes expectations about others' behavior, relevant for the first mover choice, it does not change the fact that the single player's own dominant strategy is to defect. In our experiment we exploit the equivalence between second mover BD tasks and 'second mover' responses elicited through the strategy version of the PG game.

## 2. Methods

### 2.1 Overview

The experiment was divided into two sections: a PG section (including both P- & C-games) and a BD section. Each section included sets of questions implementing the three different incentive structures introduced above. Subjects answered all questions in both sections. The order of the sections and of the question sets within each section were randomized for each subject.

### 2.2 Public goods game

The PG section included three sets of questions. In each set we implemented one of the three incentive structures outlined above. Subjects were informed that they were part of a group of 4 players. Each player was endowed with 20 points and they allocated as many of them as they wished to the public pot.<sup>7</sup> Points in the public pot were transformed into the "pot payout" according to the formula in equation (1). The pot payout was shared equally between the four group members. Each individual's payoff was therefore the total of the points they did not contribute and their share of the pot payout. The payoff for player  $i$  is

---

<sup>7</sup> Points were converted to GBP at a rate of 20 points to £1.00, for the question selected to be played out.

summarized in equation (3), where  $\pi^i$  is the payoff for player  $i$  given a contribution of  $t^i$ ,  $N$  is the total number of players, and total contributions are  $\sum_{i=1}^N t^i$ .

$$\pi^i = z^i - t^i + \frac{1}{N}(\alpha \sum_{i=1}^N t^i)^\beta \quad (3)$$

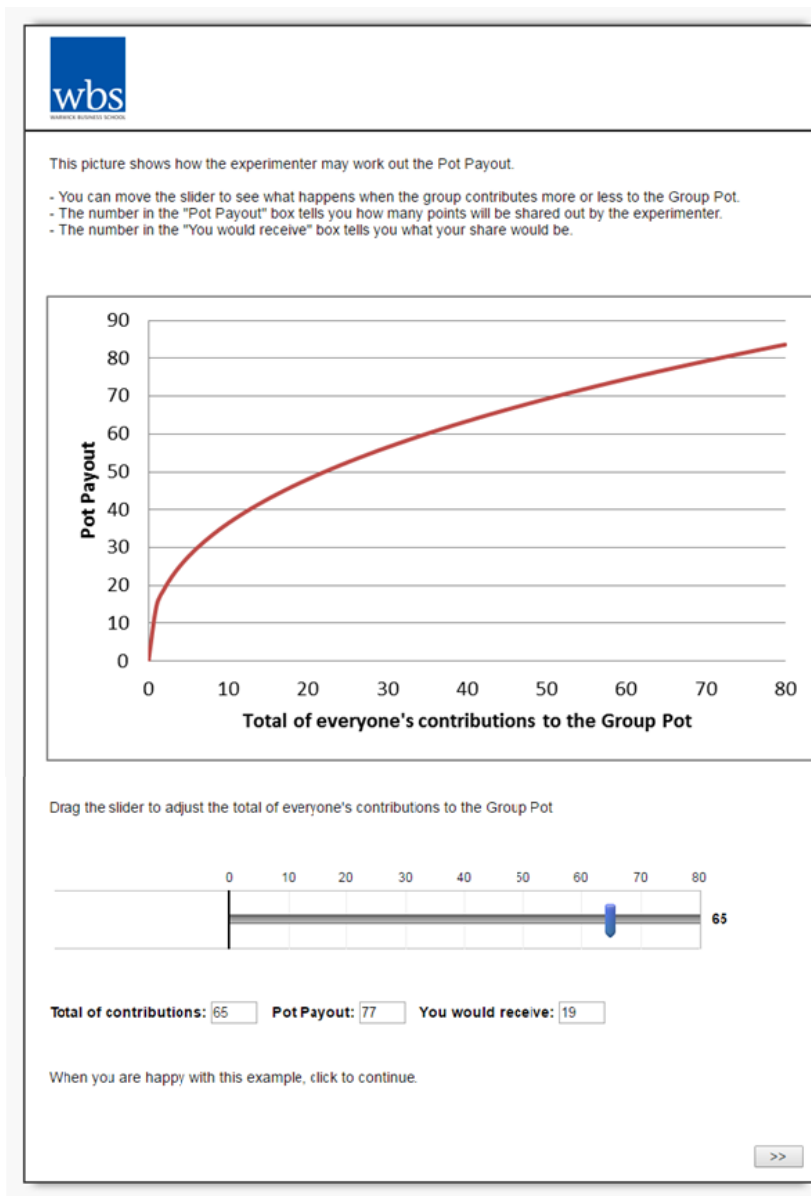
As described previously, the pot payout was a linear, concave or convex transformation of summed contributions as determined by the parameters  $\alpha$  and  $\beta$ . In the linear case  $\alpha = 1.4$  and  $\beta = 1$ ; in the convex case  $\alpha = 0.07$  and  $\beta = 3$ ; and in the concave case  $\alpha = 800$  and  $\beta = 0.4$ .<sup>8</sup>

To communicate the incentive structures in an intuitive way we developed an interactive computer interface in which the relationship between the groups' contributions to the pot and the pot payout was plotted graphically, as shown in Figure 4. The subject could use a slider to change the level of summed contributions to the pot, and the program reported the corresponding pot payout, as well as the respondent's own share of the pot payout. The amounts updated in real time as the slider was moved along. We ensured that respondents familiarized themselves with the slider by requiring them to fill in worked examples for each of the three incentive structures. We measured whether subjects understood the task by requiring them to answer a number of example questions before progressing to the main task.

FIGURE 4

---

<sup>8</sup> These parameters were chosen to produce the same incentive structure in the PG and BD tasks as outlined above (Figure 1).



**Figure 4** An example of the experiment interface for the PG task with the concave incentive structure plotted graphically. Moving the slider would update the numerical values displayed underneath in real time.

Next, subjects completed three sets of questions (one for each incentive structure). Each set included a P-game (Fischbacher et al., 2012), in which subjects were asked how many points they would contribute if they knew that the other players were going to contribute an average of  $X$  points. Responses were elicited for all possible integer values of  $X$  (0 through 20). These form the 'second mover' responses used for the main analyses. Subjects then completed a C-game, reporting how many points they would contribute if they did not know how many the other players were going to contribute. Finally (following, e.g., Fischbacher et al., 2012), subjects were asked what they expected the other player(s)' average first mover contribution to be.

## 2.3 Binary games

The binary games were based on four incentive structures: the three structures outlined above plus an incentive structure with transparent dominance and no conflict between individual and collective welfare. The latter was included as a check that respondents were paying attention and understood the task. For each of the four incentive structures subjects provided first mover responses against one other person (Figure 2), first mover responses against three other people (Figure 3), second mover responses against one other person and second mover responses against three other people. When playing against three others, subjects were told that the three other people would all independently choose their favored response and that the majority choice would be played.

FIGURE 5

Your Points Their Points	Other Chooses Left		Other Chooses Right	
You Choose Top	5		6	
	5		9	
You Choose Bottom	8		10	
	8		1	

**Figure 5** Binary choice as displayed to subjects (example, second mover), with “your points” displayed using green, and “their points” displayed using blue. This example is from the information screen and these specific payoffs were not used in any of the real tasks.

The dilemmas were presented in a grid, with different colors used to represent the payoffs for self and other(s). This presentation was supplemented by a written description of what the payoffs to the respondent and to the other player(s) would be for all four potential outcomes in that dilemma. There were two information screens: one for the choices where the subject faced one other player, and one for the choices where the subject faced three other players. After each information screen, the subject made their choices (as first and

second mover under each of the four incentive structures<sup>9</sup>). For the first mover choice, the subject chose between Top (cooperate) and Bottom (defect). The second mover choice was also presented as a decision between Top and Bottom, but this time in the knowledge that the other player(s) would choose Left. To make this clear, the irrelevant column of the table was covered with a translucent grey overlay (Figure 5). The order of blocks (1 other then 3 other or *vice versa*) was randomized between subjects, and the order of the four incentive structures within each of these blocks was also randomized. Each first mover question was always immediately followed by the relevant second mover question.

The task was incentivized by informing players that one question would be selected at random for each individual. This could be any of the choices from the first or second mover responses from either the PG or the BD tasks. If a PG task was selected, three other subjects were chosen at random to provide the “others” responses. If the subject’s role was as first mover, then their unconditional estimate was averaged with those of two others to generate the average contribution which was then rounded. The fourth player’s conditional contribution was used to complete the public pot. If their role in the PG task was second mover, the (rounded) average of the other three players’ unconditional contributions was used, and the subject’s relevant conditional cooperation amount was used to complete the public pot.

If a BD task was selected with 1 other player and the role of second mover was selected there was no need to use the responses of any other player, and the subject simply received the relevant payout. If in this task, the role of first mover was selected, then another subject was chosen at random and their first mover response was used as the other player choice. If a BD task was selected with 3 other players, then this could not be properly incentivized using subjects from within this task, as none completed a voting task (which forms the responses of the others in the 3-other condition). Therefore, an auxiliary study was conducted where subjects were told they were one of three individuals voting for the BD choice<sup>10</sup>. Three subjects from this auxiliary task were selected to provide the votes in the BD with 3 others.

---

<sup>9</sup> Although we refer to the choices as first and second mover, this does not imply that the choices were all sequential, with one person always choosing first then the second choosing based upon their response. These terms were never used in the task instructions or procedure. Subjects were told the two cases represented situations where no players knew what the other(s) would choose, and times where they already knew what the other player(s) would do.

<sup>10</sup> A total of 11 new subjects were recruited from the same prolific academic subject pool as the main study. These subjects were told that at the end of the study a random number between 1 and 10 would be selected. If the number was 10 then one of the questions would be picked and played for real. No subject got a number of 10.

The experiments were programmed online using Qualtrics, and JavaScript was used for the interactive displays. All 117 subjects in the main experiment, and the additional 11 subjects filling the role of the “others” in the binary dilemma section, were recruited through Prolific Academic and completed the experiment over the internet. Ethical approval was granted by the University of Warwick Humanities & Social Sciences Research Ethics Committee. Subjects in the main experiment received a fee of £6 for participating in addition to earnings based on their choices in the experiment. Points were converted into money at a rate of 5 pence (0.05 GBP) per point, and subjects’ choice-dependent earnings ranged from £0 to £4.50. Payment amounts were calculated and paid as soon as all data collection was complete. The mean earnings were £7.46 in total. Subjects in the short auxiliary experiment were paid £1.50 for participating in addition to having a 1 in 10 chance of receiving a reward based on their choices.

## 3. Results

### 3.1 Exclusions

We report the full results here using responses from all subjects who provided a response for every question (as a result, only three subjects were excluded). There are of course many reasonable criteria upon which one could exclude subjects from the analysis. We take a default approach of reporting analyses using the full set to ensure maximum openness and prevent any danger of selective exclusions driving significant results or of particular criteria potentially excluding subjects employing unusual strategies. However, for all of the reported analyses a parallel analysis has been performed upon a subset of subjects selected using deliberately conservative inclusion criteria. An individual was excluded from this subset if they chose the dominated option in any of the 4 binary dilemmas with the transparent dominance incentive structure, or if they began the main experiment without having answered all practice questions correctly. This resulted in 37 subjects being excluded. The qualitative result remains the same in all analyses, and quantitative differences are also small (e.g. 26% overall preference reversal, as opposed to 28% in the full sample).

### 3.2 Categorizing individuals by “traditional types” in the PG game

We begin by replicating Fischbacher et al.’s (2012) categorization, which we will refer to as the “traditional type” categorization. This method relies solely upon responses in the PG game with a linear incentive structure. Individuals are categorized as free-riders if they contribute zero regardless of the average contribution of others (13 subjects, 11%). They are

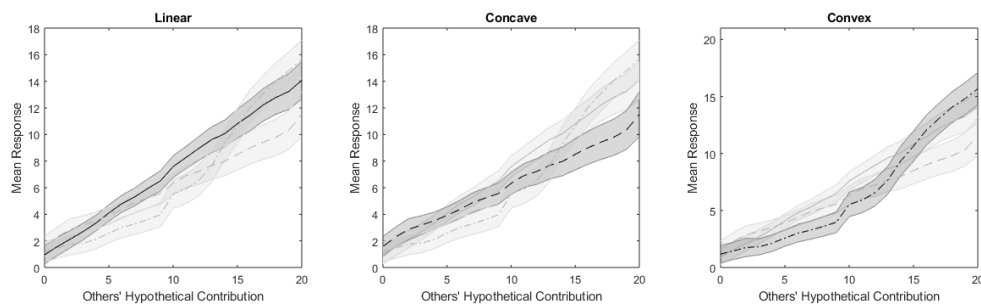


categorized as conditional cooperators if their contributions rise monotonically with others' contributions, or if there is a significant positive correlation between their contributions and the contributions of others (89 subjects, 78%). They are categorized as "triangle responders" if the maximum amount they contributed is not given in response to the maximum contributions by the other players and either a) subjects' contributions rise monotonically towards their maximum contribution and decline monotonically after that, or b) there is a significant positive correlation between their contributions and those of the other players up to the maximum point and a negative correlation thereafter (10 subjects, 9%). All others are categorized as "other" (2 subjects, 2%). These proportions are similar to those found in previous studies. Due to the low number of triangle and other responders, we follow the approach of several previous papers by combining them.

### 3.3 Responsiveness to incentive structures

We next test whether behavior differed between the incentive structures. To provide a summary view, the mean contribution and the 95% confidence intervals were calculated for each level of "others' contribution" for each incentive structure in the P-game. The results are plotted in Figure 6 which shows that, although all three patterns approximate conditional contribution, there are significant differences between the incentive structures. When "others' contributions" are low, subjects generally contribute less in the convex condition than in the concave condition. When "others' contributions" are high, subjects contribute more in the convex than in the concave incentive structure. These differences are in the direction of payoff maximization, indicating that subjects understood the incentive structures. In fact, despite the largest group being "conditional co-operators" only 9 subjects (8%) exhibited pure contribution matching across all incentive structures, whilst only 5 subjects (4%) exhibited pure non-investment. The majority of subjects (all but 14) changed their behavior in response to the incentive structure.

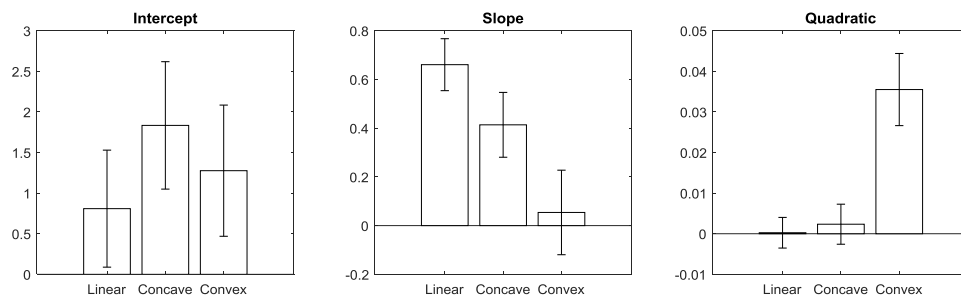
FIGURE 6



**Figure 6** Levels of contribution in the PG game under each incentive structure. The shaded area represents 95% confidence intervals around the mean.

To quantify behavior within each incentive structure, polynomial fits were estimated for each individual's set of 21 responses under each incentive structure. Second order polynomial fits were estimated. The mean coefficients and 95% confidence intervals are shown in Figure 7. The confidence intervals show that there are clear differences between incentive structures. The quadratic component is largest for responses in the convex condition, whereas the linear slope component is largest for the linear structure, and smallest for the convex.

**FIGURE 7**

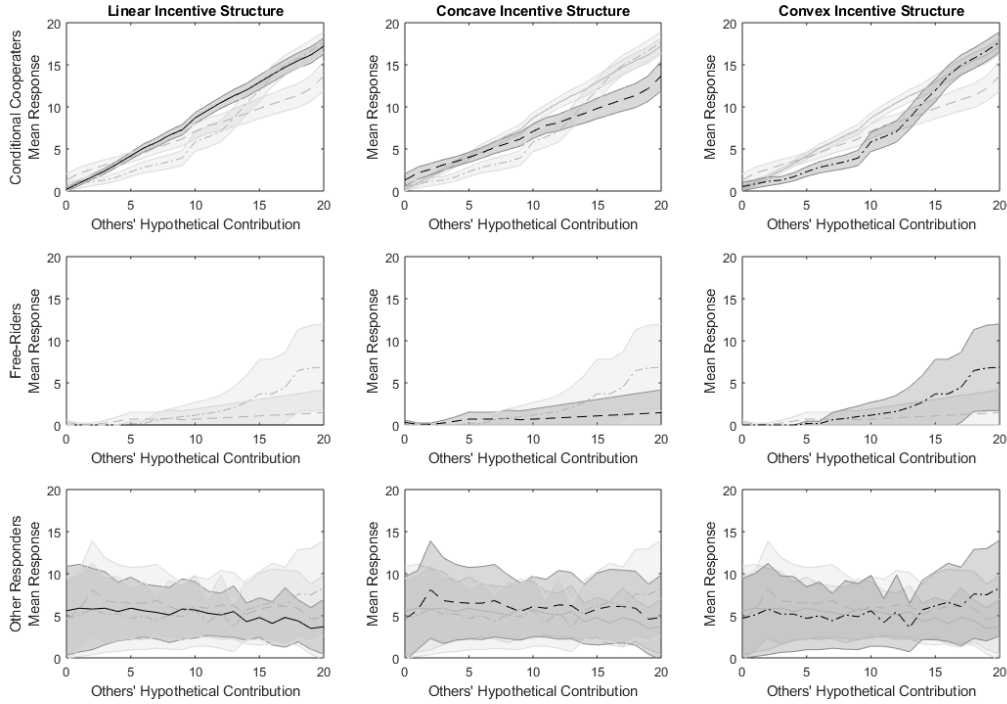


**Figure 7** Parameters of the model fits for PG game responses, by incentive structure, pooling all respondents, with 95% confidence intervals.

### 3.4 PG responses by sub-types

To examine how conditional cooperators, free-riders and triangle/other responders behaved under each incentive structure, the mean contributions and confidence intervals were plotted separately for each sub-type. Figure 8 shows clear differences between the categories. The leftmost panel shows that, despite their categorization (a strict interpretation of which would suggest they will simply match or respond to the contributions of others regardless of associated outcomes) conditional cooperators actually moderate their behavior according to the incentive structure.

**FIGURE 8**



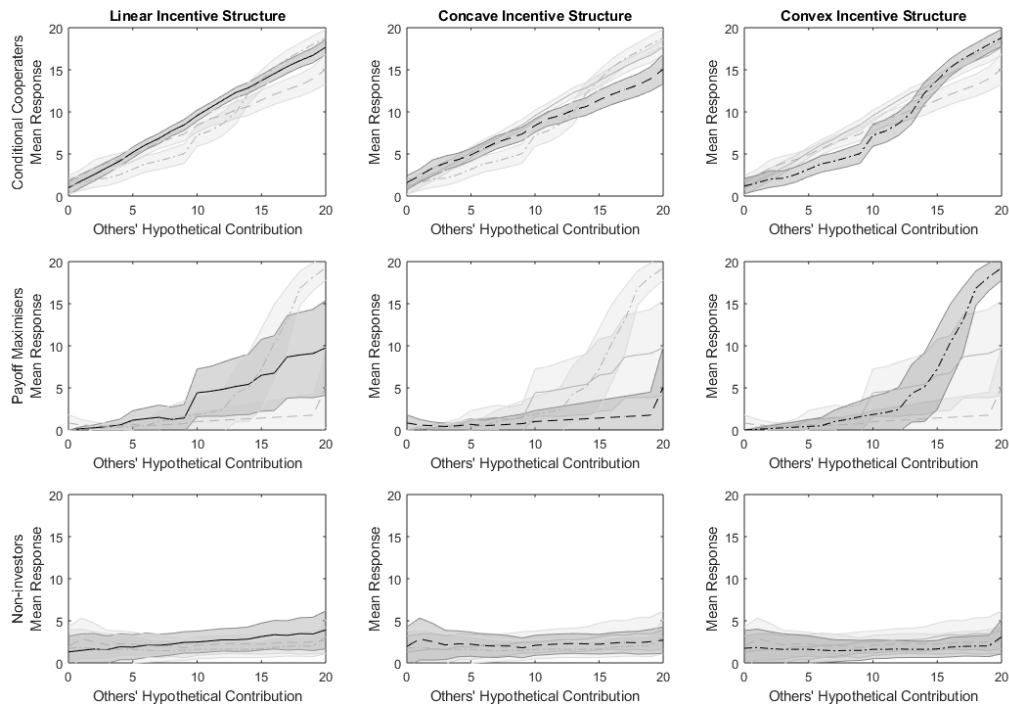
**Figure 8** Levels of contribution in the PG game in different incentive conditions with subjects categorized by the traditional Fischbacher et al., (2012) methods. The shaded areas represent 95% confidence intervals around the mean.

The middle panel illustrates a limitation of the traditional classification. Free-riders were so classified because they contributed nothing in the linear incentive structure. However, clearly they do not all exhibit this behavior in the convex or concave incentive structures. Consider for example the convex structure. Inclusion of this structure allows us to distinguish between behavioral subtypes that respond identically in the linear case. Specifically, a non-investor and a payoff-maximizer would both contribute 0 (i.e., free ride in the linear condition). In the convex structure, however, a payoff-maximizer would contribute 20, whilst a non-investor would contribute 0 in response to the maximum possible contribution of others. The large confidence intervals around the mean in the middle row panels hide an underlying bimodal distribution.

We therefore develop an alternative categorization technique based on responses across all incentive structures. The three categories are conditional cooperator, payoff-maximizer and non-investor. We first create the profile of a hypothetical exemplar responder for each type. The exemplar conditional cooperator always matches the contributions of the “other players”, the exemplar payoff-maximizer always makes the contribution that maximizes their own payoff, and the exemplar non-investor always contributes zero. Subjects are categorized as the subtype whose exemplar they most closely resemble

(resemblance is quantified in terms of summed squared distance between subjects' and exemplars' responses). As with the traditional classification method the majority of subjects were classified as conditional cooperators (79 subjects, 69%). However, our novel exemplar-based technique allowed us to classify the remaining subjects as either payoff-maximizers (12 subjects, 11%) or non-investors (23 subjects, 20%). This classification could not be achieved using the standard linear PG game. Figure 9 shows that the contribution patterns of these groups match the description of their behavior and strategy.

FIGURE 9



**Figure 9** Levels of contribution in the PG game in different incentive conditions with subjects categorized by the novel exemplar matching method. The shaded areas represent 95% confidence intervals around the mean.

### 3.5 Binary dilemmas: variability

Next we turn to BDs. Table 1 gives the proportion of subjects defecting in each BD. Much heterogeneity is evident between subjects, with most defection rates between 45% and 74%. The exceptions are the second mover choices in the stag hunt (convex structure). The very high levels of cooperation occur because once the subject knows that others will

cooperate, cooperation option dominates defection<sup>11</sup> (here we define dominating as both own and others' payoffs being higher). When comparing defection rates in first and second mover choices, there is a tendency for decreased defection rates in the prisoner's dilemma (linear structure) when acting as the second mover. In the hawk-dove game (concave structure), there is a slight tendency to increase defection, but these changes are small. For full statistical analysis see appendix B.

TABLE 1

	Facing 1 Other		Facing 3 Others	
	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover
Linear	66%	54%	74%	63%
Concave	54%	64%	62%	67%
Convex	45%	10%	55%	14%

**Table 1** The proportion of subjects choosing to defect in each binary dilemma, as 1st and 2nd mover.

### 3.6 Binary dilemmas and responder types

The overall levels of defection and cooperation may mask more homogenous patterns of behavior within groups of responders. Therefore, defection rates are shown separately for each responder sub-type; first by traditional types in Table 2 and then by our exemplar based types in Table 3. If individuals' strategies are stable across tasks, then there should be less heterogeneity within sub-types than in the sample overall. For specific sub-types there are clear predictions. For example, one would expect all free-riders and payoff maximizers to defect in all prisoner's dilemmas, and one would expect all conditional cooperators to cooperate in all second mover prisoner's dilemmas. Table 2 shows that patterns of defection rates generally match these intuitive predictions given the sub-types defined: Conditional cooperators appear generally less likely to defect than other individuals, and free-riders are generally most likely to defect. However, the effect of responder type is far from deterministic: The largest difference between types within any single question is a 48 percentage point difference in defection rates. This is the difference between the

<sup>11</sup> Note that the detected rate of defection in the second mover convex BD is the one result that shows a meaningful difference between the full and restricted sample. After excluding subjects for failure to answer attention check BDs and failing to complete understanding questions, the rates of defection fall to 1% when facing 1 other and 1% when facing 3 others.

conditional cooperators and the payoff maximizers (as defined by our exemplar method) when making second mover choices in the prisoner's dilemma.

Performing statistical analyses upon defection rates is difficult due to the small number of subjects in some groups (e.g. free riders and other) and the lack of variation in convex choices (with almost all subjects cooperating when acting as second mover). Logistic regressions were performed for each combination of first/second mover and number of other players, separately for the traditional types and exemplar types. Full results are presented in Appendix B but, to summarize, the only comparison showing any significant main effect of responder type is when subjects were acting as the second mover against 1 other player. The traditionally-defined free-riders, and the exemplar-defined payoff maximizers and non-investors were all more likely to defect than the corresponding conditional cooperators. However, comparing the log-likelihoods of different models demonstrates that the exemplar method is more accurate at predicting second mover responses.

TABLE 2

	<b>Linear</b>		<b>Concave</b>		<b>Convex</b>	
<b>Facing 1 Other</b>	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover
<b>Conditional</b>	65%	51%	53%	58%	44%	8%
<b>Free-Rider</b>	85%	85%	77%	85%	62%	15%
<b>Other</b>	50%	60%	30%	90%	40%	20%
<b>Facing 3 Others</b>						
<b>Conditional</b>	72%	58%	61%	62%	56%	12%
<b>Free-Rider</b>	92%	100%	77%	92%	62%	15%
<b>Other</b>	80%	70%	70%	90%	50%	30%

**Table 2** Proportion of subjects choosing to defect in each binary dilemma; subjects are categorized by the traditional Fischbacher et al., (2012) method.

TABLE 3

	<b>Linear</b>		<b>Concave</b>		<b>Convex</b>	
<b>Facing 1 Other</b>	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover	1 <sup>st</sup> Mover	2 <sup>nd</sup> Mover
<b>Conditional</b>	62%	44%	49%	58%	43%	8%

<b>Payoff Maximizers</b>	83%	92%	67%	92%	42%	8%
<b>Non-Investors</b>	70%	70%	61%	70%	52%	17%

#### **Facing 3 Others**

<b>Conditional</b>	71%	57%	59%	62%	53%	14%
<b>Payoff Maximizers</b>	83%	83%	75%	83%	67%	8%
<b>Non-Investors</b>	78%	74%	65%	74%	57%	17%

**Table 3** Proportion of subjects choosing to defect in each binary dilemma; subjects are categorized by the novel exemplar matching method.

### **3.7 Preference Reversals**

To provide a more robust test of cross-task consistency, each individual's responses in one task were compared to their responses in the equivalent incentive structure of the other task to identify preference reversals. To demonstrate how preference reversals are identified, take the linear structure as an example. The choice to defect in the BD is equivalent to contributing 5 in the PG. The choice to cooperate in the BD is equivalent to contributing 20 in the PG. Therefore, anyone who cooperates in the BD, and then contributes 5 in the PG exhibits an unambiguous reversal of preference. This is also true for anyone who cooperates and then contributes less than 5. For individuals who defect in the BD, a contribution of 20 in the PG indicates an unambiguous preference reversal. Since we are interested in cases of unambiguous reversals, we restrict our analysis to subjects who contribute an amount in the PD equal to or greater than the amount equivalent to cooperating in the BD, and those who contribute an amount in the PG equal to or smaller than the amount equivalent to defecting in the BD.

Table 4 shows in brackets the total number of the 114 subjects who contributed an amount greater than or less than these values, and the associated percentage indicates the proportion of those subjects who indicated the reverse preference in their BD choice. Looking across all subjects regardless of whether their level of PG contribution would make it possible to unambiguously identify preference reversals, we find that in the second mover condition with three others, 30.7% of all subjects unambiguously reversed their preference in the linear condition, 36.8% did so in the concave condition, and 15.8% in the convex condition.

**TABLE 4**

	<b>1<sup>st</sup> mover</b>		<b>2<sup>nd</sup> Mover</b>	
<b>1 Other</b>	Cooperation given <min contribution	Defection given >max contribution	Cooperation given <min contribution	Defection given >max contribution
<b>Linear</b>	29% (24)	58% (60)	21% (24)	47% (60)
<b>Concave</b>	48% (27)	53% (53)	22% (27)	53% (53)
<b>convex</b>	46% (13)	44% (77)	77% (13)	6% (77)
<b>3 Others</b>				
<b>Linear</b>	21% (24)	68% (60)	13% (24)	53% (60)
<b>Concave</b>	37% (27)	62% (53)	30% (27)	64% (53)
<b>convex</b>	38% (13)	55% (77)	85% (13)	9% (77)

**Table 4** The numbers in brackets indicate the number of subjects who in the PG game contributed at least the amount equivalent to cooperating in the BD, and who contributed an amount no more than that equivalent to defecting in the BD. The associated percentages indicate the proportion of those subjects who exhibited the reverse preference in their BD choice.

### 3.8 Predictions across tasks using continuous measures

Despite the significant proportion of preference reversals, it remains possible that there is a link between behavior in the two tasks. This link is clearly not deterministic, and not well described by a subject's categorization, but an individual who exhibits 'cooperative' response patterns in the PG game may be more likely to cooperate in the BDs. To examine this, we return to the individual-level polynomial regressions outlined earlier and use the estimated coefficients from the regressions upon behavior in the PG game (see section 3.3). A further, logistic regression was used to predict binary dilemma choices using these (z-scored) estimates of curve, slope and intercept as independent variables. All incentive structures were included in the one model on separate rows, e.g. the polynomial coefficients estimated from a subject's responses in the linear PG game were used to predict the likelihood that they would defect in the linear BD games, and those from the concave PG game were used to predict responses from the concave BD games etc.

The model intercept is -0.13 (CI = [-0.25, 0.10],  $p = 0.258$ ). The strongest predictor is the polynomial intercept term (Beta = -1.33, CI = [-1.79, -0.87],  $p < 0.001$ ), with smaller significant results for the polynomial linear term (Beta = -0.74, CI = [-1.20, -0.27],  $p = 0.002$ )



and the polynomial quadratic term (Beta = -0.39, CI = [-0.68, -0.10],  $p = 0.009$ ). All three coefficients are negative indicating that higher intercept, steeper slope and greater curvature in an individual's PG response profile predict lower probability of defecting in BDs. The model overall shows that the polynomial coefficients are significantly predictive, with a McFadden's R-Square of 0.102.

A similar analysis can be performed on the first mover responses. In the PG game, each first mover response is an integer between 0 and 20. This was entered into a logistic regression (along with an intercept term) to predict whether the individual then defected in the corresponding first mover choice in the BD game. The intercept term is larger than zero (Beta = 0.57, CI = [0.35, 0.79],  $p < 0.001$ ) indicating high baseline levels of defection. There is also a significant effect of PG first mover contribution in the predicted direction (Beta = -0.37, CI = [-0.61, -0.14],  $p = 0.002$ ), meaning that if subjects contribute more in the PG then they are less likely to defect in the BD. However, the overall accuracy of predictions (McFadden's R-Square of 0.025) is not as good as in the case of the second mover responses. This is likely due to subjects displaying caution due to uncertainty about other players' contribution levels, and the additional variability caused by different subjects making different assumptions about the other players' likely contribution levels.

### **3.9 Generalized individual-level traits or structure-specific strategies?**

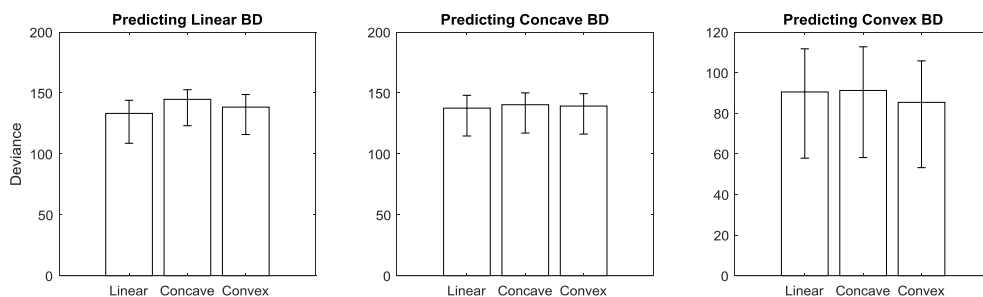
Finally, we ask whether individuals can be described as having an underlying propensity to cooperate in all situations. Such a tendency would be akin to a personality trait, which could be labelled "pro-sociality". If so, cross-task predictability should be as good across as within incentive structures. The alternative possibility is that individuals' cooperativeness is specific to incentive structures. If this is so, individuals' rates of cooperation will show greater cross-task predictability when the incentive structures are identical.

For this analysis, we return to the individual-level polynomial regressions outlined earlier and use the estimated coefficients from the regressions upon behavior in the PG game. To examine whether strategies are specific to incentive structures, a logistic regression was used to predict binary dilemma choices using these estimates of curve, slope and intercept.

Specifically, we examined whether an individual's cooperation in the BDs under one incentive structure is better predicted by the estimated parameters obtained from their responses in the PG game within that incentive structure than by the parameters obtained from the PG games with the other incentive structures. Therefore, 9 separate regression

models were estimated – one for every pairwise combination of PG incentive structure and BD incentive structure. The deviance (a goodness of fit measure) was then used to assess the predictive accuracy of each of the 9 models. To provide confidence intervals, a bootstrapping approach was used. For each regression, 10,000 bootstrap samples were randomly selected with replacement from the subjects in our dataset. Figure 10 shows the deviance estimates for each model and the confidence intervals around them. The deviance is very similar regardless of the incentive structure used to estimate the PG parameters, and the confidence intervals overlap substantially. Thus there is no evidence of any advantage to using incentive structure-specific estimates of cooperativeness, consistent with the idea that pro-sociality is a trait.

FIGURE 10



**Figure 10** Deviance estimates for each model with 95% confidence intervals, demonstrating no improvement in prediction within than between incentive structures. Note that the asymmetrical confidence intervals are a natural result of the bootstrap estimation method: for some models there is a negative skew in the deviance of the bootstrap samples.

## 4. Discussion

The viability of predicting behavior in laboratory social dilemmas from behavior on other laboratory tasks has not been well-established, despite a number of investigations (Apesteguia & Maier-Rigaud, 2006; Blanco et al., 2007; Binmore & Shaked, 2010; Camerer, 2011; Galizzi & Navarro-Martinez, 2018; Kingsley & Liu, 2014). If such a link cannot be established, claims that results in the laboratory can inform us about behavior in decision contexts outside the laboratory where the stakes, framing and incentives all differ are premature. The results presented here established just such cross-task predictability, but also note its limitations. More specifically, we find that contribution levels in PG tasks predict choices in BD tasks, but that there is significant stochasticity and that even strictly defined responder types do not deterministically predict responses in different tasks, even when the incentives are identical.

We obtained our results from what is, to our knowledge, the first experiment that compares behavior within subjects across PG games and binary social dilemmas with identical financial incentives. Our novel paradigm allowed us to test the validity and generalizability of laboratory social dilemmas more strictly than has been previously possible, since all normatively relevant aspects of the decisions can be held constant. We addressed two key questions. First we asked what strategies were employed in the PG game when incentive structures were modified, including whether the traditionally identified categories of individuals could also characterize behavior in non-linear incentive structures. Second we asked whether behavior in the PG games predicts behavior in BDs, and whether this predictive ability is specific to particular incentive structures or if strategies predict behavior across incentive structures as well as across tasks.

We applied the traditional categorization of individuals according to their behavior in the PG game with a standard, linear incentive structure (Fischbacher et al., 2012), and found proportions of conditional cooperators, free-riders and other responders that were in keeping with the existing literature. However, the standard categorization based on the linear PG game confounds own-payoff maximization and a simple strategy of non-investment, since these produce indistinguishable behavior (i.e., non-contribution at all levels). By introducing different incentive structures and categorizing individuals by their fit with “exemplar” subtypes, we were able to distinguish payoff maximization and non-investment. We find that many individuals who are traditionally categorized as free-riders can more accurately be described as either payoff-maximizers or non-contributors. This alternative categorization better captures the contributions in the non-linear incentive structures, suggesting that the traditional categorization method cannot be extended beyond the linear PG task.

Both the non-contributors and the payoff-maximizers were sensitive to changes in the incentive structure, with their aggregate behavior adapting to changing incentives in the direction that would improve their own payoff. This behavior is an important departure from that predicted by a strict interpretation of conditional cooperation, according to which individuals match the contributions of others. Instead, this result suggests that their responses reflect a compromise between pro-social preferences for contribution or outcome equality, and payoff maximization. We see little evidence that the behavior of these individuals is the result of simple imitation strategies, with only 9% of subjects always matching the contributions of the other players.

Next we turn to the question of cross-task predictability. Subjects responded to a set of BDs as first and second mover in addition to the PG game. These BDs were designed such that the financial incentive structures matched those of the PG games at the levels of

contribution equivalent to defect and cooperate choices. This property is crucial as it allows the direct assessment of cross-task consistency whilst the incentive structure is held constant.

We found a significant relationship between an individual's likelihood of cooperating in the PG game and their BD choices. This relationship was weak when using PG game categorization to predict BD choices, with only one analysis showing a statistically significant increase in defection for payoffs maximizers. However, by using the continuous nature of PG responses to make more fine-grained probabilistic predictions of BD choices, accuracy was dramatically increased and the relationship was significant across tasks.

A further question was whether cross-task consistency only occurs when incentive structures are identical. If individuals applied qualitatively different strategies under different incentive structures then it might be possible to predict BD choices from PG contributions when incentive structures match (e.g. linear from linear), and yet not possible to predict BD choices from PG contributions when incentive structures differ (e.g. linear from convex). An alternative possibility is that cross-task prediction is possible but is not sensitive to incentive structure. Such a result would be consistent with the existence of an individual-level trait, such as pro-sociality or other-regarding preference that influences behavior in all tasks and incentive structures. The results favor the latter possibility.

Overall, there is a significant relationship between responses in the PG and BD tasks. However, this link is far from deterministic. When the rates of cooperation in the BDs were calculated separately for different PG responder types, the differences in choice proportions were in line with intuitive expectations, but there was still much heterogeneity within responder types. If subgroups were applying a particular strategy consistently then one would expect individuals of the same type to display the same behavior. To illustrate, define "high agreement" as occurring when at least 90% of individuals within a group respond in the same way on a given choice. This level of agreement was found in only 14% of cases<sup>12</sup> (excluding questions involving dominance) when subjects were classified by our exemplar method, and in 17% of cases when subjects were classified using the traditional method (see Tables 2 & 3). Furthermore, in 28% of cases, individuals exhibited unambiguous preference reversals between tasks – either by cooperating in the binary dilemma and contributing less than the defect amount in the PG game, or by defecting in the binary dilemma and contributing more than the cooperate amount in the PG game.

---

<sup>12</sup> A "case" here refers to a subgroup-question combination as defined in the cells of Tables 2 and 3.

Preference reversals in the former direction are substantially more common than in the latter.

In summary we have two findings that appear hard to reconcile: There is significant cross-task predictive power, but there is also significant within-group heterogeneity with even unambiguous preference reversals being relatively common. However, when considered as a whole, these results support the hypothesis that individuals have stable pro-sociality traits that influence their propensity to cooperate, whilst at the same time their strategies can also vary between tasks. Such variation could reflect framing and other structural differences. A stable trait such as this would act as a parameter on task-specific strategies. The presence of this same parameter value in different task strategies explains the cross-task correlations in behavior. However, the fact that individuals can be differentially sensitive to particular framing effects, and that different strategies can be applied by different individuals in each task explains the heterogeneity in responses. This interpretation also explains why behavior cannot be better predicted within an incentive structure, as individuals are likely to employ the same strategy within a task, regardless of the incentive structure associated with that specific question.

Our framework of stable traits and changing strategies may seem something of a departure from typical approaches in experimental economics. However, it is actually very similar to the approach taken in many other parts of the literature. As an analogue, consider how individuals respond in tasks involving risky financial gambles. Individuals respond significantly differently in choice tasks than in valuation tasks, similarly to how they behave differently in different social dilemma tasks. However, an individual who displays strong risk aversion when choosing between risky gambles is also likely to show strong risk aversion when valuing risky gambles. This risk aversion may manifest itself differently in the two tasks; nonetheless, once one has a robust model of behavior in each task, risk aversion can be estimated in one, and then used to predict responses in the other. Our results suggest that the same approach can be used in social dilemma tasks, opening a new route to understanding pro-social behavior in a variety of settings.

Why do we find consistent levels of cooperation across both tasks and incentive structures while a number of other studies have not? There are two possible reasons for the different patterns of findings. On one interpretation, studies which have failed to find cross-task or cross-incentive-level correlations may be theoretically misconceived in that their chosen tasks do not assess the same dimensions of pro-sociality. A second possibility is that extraneous and theoretically irrelevant / uninteresting noise variables explain the discrepancy by in some way either overwhelming or obscuring effects of a true underlying

prosociality trait. We incline towards the latter interpretation. Our study differs from earlier experiments in several ways, in particular by holding incentive structures constant across task frames. Further research will be needed to identify the specific features that determine whether evidence for a stable personality-like pro-sociality trait will be found in a particular study, but the results we have presented here, in providing evidence for cross-task prosociality under carefully controlled experimental conditions, may provide at least a starting-point for such an exploration.

# References

- Andreoni, J. (1995). Cooperation in public-goods experiments - kindness or confusion. *American Economic Review*, 85(4), 891-904.
- Apesteguia, J., & Maier-Rigaud, F. P. (2006). The role of rivalry - Public goods versus common-pool resources. *Journal of Conflict Resolution*, 50(5), 646-663.
- Binmore, K., & Shaked, A. (2010). Experimental economics: Where next? *Journal of Economic Behavior & Organization*, 73(1), 87-100.
- Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2), 321-338.
- Camerer, C. F. (2011). The promise and success of lab-field generalizability in experimental economics : A critical reply to Levitt and List \*.
- Carpenter, J., & Seki, E. (2011). Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry*, 49(2), 612–630.
- Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* 14 (1), 47-83
- Conybeare, J. A. C. (1984). Public-goods, prisoners' dilemmas and the international political-economy. *International Studies Quarterly*, 28(1), 5-22.
- Dai, Z., Galeotti, F., & Villeval, M. C. (2016). Cheating in the Lab Predicts Fraud in the Field. An Experiment in Public Transportations. Ssrn, (May).  
<http://doi.org/10.2139/ssrn.2725911>
- Fehr, E., & Leibbrandt, A. A field study on cooperativeness and impatience in the Tragedy of the Commons, 95 Journal of Public Economics 1144–1155 (2011).  
<http://doi.org/10.1016/j.jpubeco.2011.05.013>
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817-868.
- Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3), 397-404.
- Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, 33(4), 897-913.
- Hauert, C., & Szabo, G. (2003). Prisoner's dilemma and public goods games in different geometries: compulsory versus voluntary interactions. *Complexity*, 8(4), 31-38.

- Isaac, R. M., & Walker, J. M. (1988). Group-size effects in public-goods provision - the voluntary contributions mechanism. *Quarterly Journal of Economics*, 103(1), 179-199.
- Galizzi, M. M., & Navarro-Martinez, D. (2018). On the External Validity of Social Preference Games: A Systematic Lab-Field Study. *Management Science*, mns.2017.2908.
- Karlan, D. S. (2005). Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *American Economic Review*, 95(5), 1688–1699.  
<http://doi.org/10.1257/000282805775014407>
- Keser, C. (1996). Voluntary contributions to a public good when partial contribution is a dominant strategy. *Economics Letters*, 50(3), 359-366.
- Kingsley, D. C., & Liu, B. (2014). Cooperation across payoff equivalent public good and common pool resource experiments. *Journal of Behavioral and Experimental Economics*, 51, 79-84.
- Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, 24, 183-214.
- Lamba, S., & Mace, R. (2011). Demography and ecology drive variation in cooperation across human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(35), 14426–30.  
<http://doi.org/10.1073/pnas.1105186108>
- Ledyard, J. O. "Public Goods: A Survey of Experimental Research." (1994).
- Ledyard, J. O., & Palfrey, T. R. (1995). Experimental game-theory - introduction. *Games And Economic Behavior*, 10(1), 1-5.
- Normann, H. T., Requate, T., & Waichman, I. (2014). Do short-term laboratory experiments provide valid descriptions of long-term economic interactions? A study of Cournot markets. *Experimental Economics*, 17(3), 371–390. <http://doi.org/10.1007/s10683-013-9373-9>
- Sefton, M., & Steinberg, R. (1996). Reward structures in public good experiments. *Journal of Public Economics*, 61(2), 263-287.
- Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopol-experiments, S. 136–168. *Beiträge zur experimentellen Wirtschafts-Forschung*. Tübingen: JCB Mohr.
- Stoop, J., Noussair, C. N., & van Soest, D. (2012). From the Lab to the Field: Cooperation among Fishermen. *Journal of Political Economy*, 120(6), 1027–1056.  
<http://doi.org/10.1086/669253>



# Appendix A

Calculating equilibria in PG games with our incentive structures

The payoff to any individual is given as

$$\pi^i = 20 - t^i + \frac{1}{4} \left( \alpha (T^J + t^i) \right)^\beta$$

To differentiate by  $t^i$  we need the chain rule

$$\text{Let } u = \alpha (T^J + t^i)$$

Then

$$\pi^i = 20 - \frac{u}{\alpha} + T^J + \frac{1}{4} (u)^\beta$$

$$\frac{\partial \pi^i}{\partial u} = -\frac{1}{\alpha} + \frac{\beta}{4} (u)^{\beta-1} = -\frac{1}{\alpha} + \frac{\beta}{4} \left( \alpha (T^J + t^i) \right)^{\beta-1}$$

$$\frac{\partial u}{\partial t^i} = \alpha$$

$$\frac{\partial \pi^i}{\partial t^i} = \frac{\partial \pi^i}{\partial u} \frac{\partial u}{\partial t^i} = -1 + \frac{\alpha \beta}{4} \left( \alpha (T^J + t^i) \right)^{\beta-1}$$

So, contributing an additional point to the pot is beneficial to a self-interested individual as long as

$$1 < \frac{\alpha \beta}{4} \left( \alpha (T^J + t^i) \right)^{\beta-1}$$

## Appendix B

The marginal effects on increase defection in BD games across all subjects.

	Beta	p-Value	95% Cis
intercept	0.654***	0.001	[0.267, 1.041]
Concave	-0.513	0.06	[-1.047, 0.021]
Convex	-0.865**	0.002	[-1.400, -0.330]
3-others	0.376	0.195	[-0.193, 0.944]
2 <sup>nd</sup> mover	-0.478	0.08	[-1.012, 0.056]
Concave * 3-others	-0.015	0.97	[-0.791, 0.761]
Concave * 2 <sup>nd</sup> mover	0.914*	0.017	[0.161, 1.668]
Convex * 3-others	0.047	0.905	[-0.725, 0.819]
Convex * 2 <sup>nd</sup> mover	-1.547***	0.001	[-2.447, -0.648]
3-others * 2 <sup>nd</sup> mover	-0.013	0.975	[-0.790, 0.765]
Concave * 3-others * 2 <sup>nd</sup> mover	-0.232	0.676	[-1.319, 0.855]
Convex * 3-others * 2 <sup>nd</sup> mover	0.014	0.982	[-1.228, 1.256]
LogLikelihood	-833.0		

The marginal effects on increase defection in BD games when acting as the first mover and playing against one other player. Predictors are the PG category defined using the Fashbacher et al. method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	0.63**	0.005	[0.19, 1.06]
Free-Riders	1.08	0.178	[-0.49, 2.65]
Other	-0.63	0.311	[-1.84, 0.59]
Concave	-0.51	0.095	[-1.12, 0.09]
Convex	-0.87**	0.005	[-1.48, -0.27]
Free-Riders*Concave	0.01	0.99	[-2.06, 2.09]
Other*Concave	-0.36	0.72	[-2.33, 1.61]
Free-Riders*Convex	-0.18	0.842	[-1.94, 1.58]
Other* Convex	0.18	0.839	[-1.58, 1.94]
LogLikelihood	-449.9		

The marginal effects on increase defection in BD games when acting as the first mover and playing against three other players. Predictors are the PG category defined using the Fashbacher et al. method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	0.94***	<0.001	[0.48, 1.4]
Free-Riders	1.54	0.148	[-0.55, 3.64]
Other	-0.25	0.707	[-1.53, 1.04]
Concave	-0.51	0.114	[-1.13, 0.12]
Convex	-0.69*	0.03	[-1.32, -0.07]
Free-Riders*Concave	-0.77	0.543	[-3.27, 1.72]
Other*Concave	-1.32	0.281	[-3.73, 1.08]
Free-Riders*Convex	0.15	0.869	[-1.63, 1.93]
Other* Convex	-0.34	0.709	[-2.11, 1.44]
LogLikelihood	-433.3		

The marginal effects on increase defection in BD games when acting as the second mover and playing against one other player. Predictors are the PG category defined using the Fashbacher et al. method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	0.02	0.916	[-0.39, 0.44]
Free-Riders	1.68*	0.035	[0.12, 3.25]
Other	-0.02	0.971	[-1.23, 1.18]
Concave	0.32	0.293	[-0.27, 0.91]
Convex	-2.48***	<0.001	[-3.36, -1.61]
Free-Riders*Concave	-0.32	0.778	[-2.53, 1.89]
Other*Concave	-0.93	0.431	[-3.23, 1.38]
Free-Riders*Convex	1.29	0.202	[-0.69, 3.28]
Other* Convex	0.87	0.412	[-1.21, 2.96]
LogLikelihood	-365.0		

The marginal effects on increase defection in BD games when acting as the second mover and playing against three other players. Predictors are the PG category defined using the Fashbacher et al. method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	0.34	0.114	[-0.08, 0.76]
Free-Riders	113.28	1	[-3.6e+7, 3.6e+7]
Other	0.00	0.995	[-1.23, 1.22]
Concave	0.14	0.646	[-0.46, 0.74]
Convex	-2.30***	<0.001	[-3.06, -1.54]
Free-Riders*Concave	-111.27	1	[-3.6e+7, 3.6e+7]
Other*Concave	-113.02	1	[-3.6e+7, 3.6e+7]
Free-Riders*Convex	0.62	0.508	[-1.22, 2.46]

Other* Convex	0.86	0.372	[-1.03, 2.76]
LogLikelihood	-367.3		

Predicting defection in BD games when acting as the first mover and playing against one other player. Predictors are the PG category defined using the exemplar method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	0.49*	0.034	[0.04, 0.94]
Payoff-Maximizers	1.12	0.166	[-0.47, 2.7]
Non-Investor	0.34	0.509	[-0.66, 1.33]
Concave	-0.52	0.110	[-1.15, 0.12]
Convex	-0.77*	0.018	[-1.41, -0.13]
Payoff-Maximizers *Concave	-0.4	0.700	[-2.44, 1.64]
Non-Investor*Concave	-1.17	0.251	[-3.18, 0.83]
Payoff-Maximizers *Convex	0.13	0.852	[-1.24, 1.51]
Non-Investor* Convex	0.03	0.964	[-1.33, 1.4]
LogLikelihood	-455.7		

Predicting defection in BD games when acting as the first mover and playing against three other players. Predictors are the PG category defined using the exemplar method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	0.89***	<0.001	[0.40, 1.38]
Payoff-Maximizers	0.72	0.376	[-0.87, 2.31]
Non-Investor	0.39	0.487	[-0.71, 1.49]
Concave	-0.51	0.134	[-1.17, 0.16]
Convex	-0.76*	0.023	[-1.42, -0.11]
Payoff-Maximizers *Concave	-0.01	0.996	[-2.11, 2.1]
Non-Investor*Concave	-0.15	0.883	[-2.20, 1.89]
Payoff-Maximizers *Convex	-0.15	0.845	[-1.62, 1.32]
Non-Investor* Convex	-0.26	0.729	[-1.70, 1.19]
LogLikelihood	-436.0		

Predicting defection in BD games when acting as the second mover and playing against one other player. Predictors are the PG category defined using the exemplar method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	-0.23	0.312	[-0.67, 0.22]
Payoff-Maximizers	2.63*	0.014	[0.53, 4.72]
Non-Investor	1.06*	0.037	[0.06, 2.05]
Concave	0.56	0.081	[-0.07, 1.19]
Convex	-2.27***	<0.001	[-3.21, -1.33]
Payoff-Maximizers *Concave	-0.56	0.711	[-3.52, 2.4]
Non-Investor*Concave	-2.53	0.104	[-5.57, 0.52]
Payoff-Maximizers *Convex	-0.56	0.434	[-1.97, 0.84]
Non-Investor*Convex	-0.11	0.894	[-1.8, 1.57]
LogLikelihood	-356.8		

Predicting defection in BD games when acting as the second mover and playing against three other players. Predictors are the PG category defined using the exemplar method, and the type of BD game being played.

	Beta	p-Value	95% Cis
intercept	0.28	0.217	[-0.17, 0.73]
Payoff-Maximizers	1.33	0.1	[-0.25, 2.91]
Non-Investor	0.76	0.148	[-0.27, 1.79]
Concave	0.21	0.517	[-0.43, 0.85]
Convex	-2.1***	<0.001	[-2.88, -1.32]
Payoff-Maximizers *Concave	-0.21	0.854	[-2.45, 2.03]
Non-Investor*Concave	-1.91	0.161	[-4.57, 0.76]
Payoff-Maximizers *Convex	-0.21	0.778	[-1.67, 1.25]
Non-Investor*Convex	-0.5	0.548	[-2.12, 1.12]
LogLikelihood	-379.2		